A Guide to Biostatistics in Clinical Research

Dear Reader,

Biostatistics is crucial for obtaining and interpreting results from clinical data to assess the safety and efficacy of a drug, treatment, or therapy. This e-book is intended for early-stage clinical researchers and students who are interested in learning about use of biostatistics in clinical studies and are familiar with certain basic statistical concepts. The information in this e-book will also help readers who are interested in understanding the statistical approach and methods used in scientific research papers. In this e-book, we have attempted to compile some of the essential information related to topics such as hypothesis testing, error-types, power, and sample size.

We have also provided some popular additional resources at the end of the e-books for your reference. These pieces will also lead you to the original sources, which will definitely provide you detailed information.

Happy Reading!

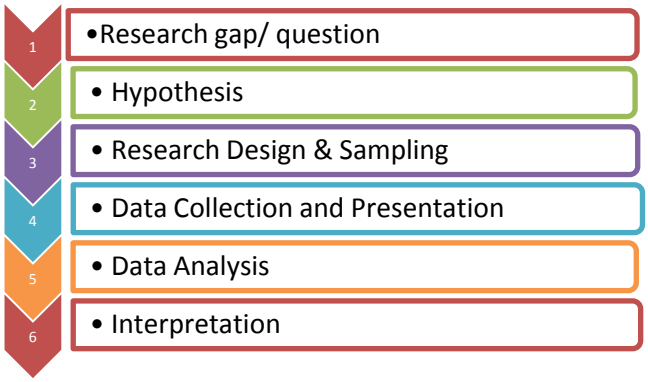
Copyright Notice | All content used in this ebook is owned or licensed by Crimson Interactive Inc. or its affiliates under the CC BY-NC-SA 4.0 license. Unauthorized use of any part of this ebook by any other party is prohibited.

Contents

1. Hypothesis 3
1.1 What Is A Hypothesis? 3
1.2 Different Types of Hypothesis 4
1.3 Study Designs 5
2. Hypothesis Testing 6
2.1 Different Types of Hypothesis Tests 6
2.2 P-Value Approach 7
2.3 Misinterpretation of P-Value 7
3. Types of Errors 9
3.1 Type I (False Positive) 9
3.2 Type II (False Negative) 9
4. Power of a Statistical Test 10
4.1 Effect Size 10
4.2 What Factors Affect Power? 11
5. Sample Size 12
5.1 What Factors Affect Sample Size? 12
5.2 How to Calculate Sample Size? 12
6. Relative Risk & Odds Ratio 13
6.1 Odds Ratio (OR) 13
6.2 Relative Risk (RR) 13
7. Correlation & Regression 14
7.1 Regression Analysis 14
7.2 Correlation 15
7.3 Types of Correlation 15
7.4 Note of Caution 17
8. Multiplicity Issue 18
8.1 Multiplicity Adjustment 18
8.2 Clinical Trial Adjustments 19

1. Hypothesis

Any research idea starts with identifying a gap in the present knowledge, i.e. framing a research question. This step is followed by proposing a hypothesis, designing a study to test the hypothesis, collecting and presenting research data, analysing the data, and interpreting the results obtained. In medical research, evidence is collected by conducting clinical trials in various phases. Clinicians arrive at the conclusion regarding safety and efficacy of the proposed drug, line of treatment, or therapy after analysing the research data.



1.1 What Is A Hypothesis?

A hypothesis is the starting point of a clinical study. It is defined as a statement that describes the relationship between two or more variables and can be proven or disproven by supporting data. Hence, a hypothesis can anticipate about a population parameter or relationships among population parameters.

A simple hypothesis consists of a predictor and an outcome variable. However, if the hypothesis consists of multiple variables, then it should be broken down to avoid use of multiple statistical tests.

Characteristics of a good hypothesis include:

1. Simplicity
2. Clarity
3. Impartiality
4. Specificity
5. Objectivity
6. Relevance
7. Verifiability

1.2 Different Types of Hypothesis

There are two types of hypothesis:

Null hypothesis (H_0): It states that there is no relationship between the predictor and the outcome variable in the population studied. It is assumed true but evidence is gathered to disprove this prediction. There is an association between the consumption of 200 g of chocolate/day and reduced risk of heart attack.

Alternative hypothesis (H_1): It states that there is a relationship between the predictor and the outcome. Clinicians try to prove this prediction. There is an association between high birth weight and obesity during adulthood.

It is impossible to prove a statement by making several observations; however, it is possible to disprove a statement with a single observation, which is why we test the null hypothesis. This is also the reason why an alternative hypothesis cannot be directly tested.

The alternative hypothesis proposed in medical research can be either **one-tailed** or **two-tailed**.

A one-tailed alternative hypothesis would predict the direction of the effect. For instance, clinical studies may have an alternative hypothesis that patients taking the drug used in the study will have a lower cholesterol level than those taking a placebo.

A two-tailed alternative hypothesis would only state that there is an association without specifying the direction. For instance, patients who take the drug used in the study will have significantly different cholesterol levels than those patients taking a placebo. The alternative hypothesis does not state whether that level will be higher or lower than those taking the placebo.

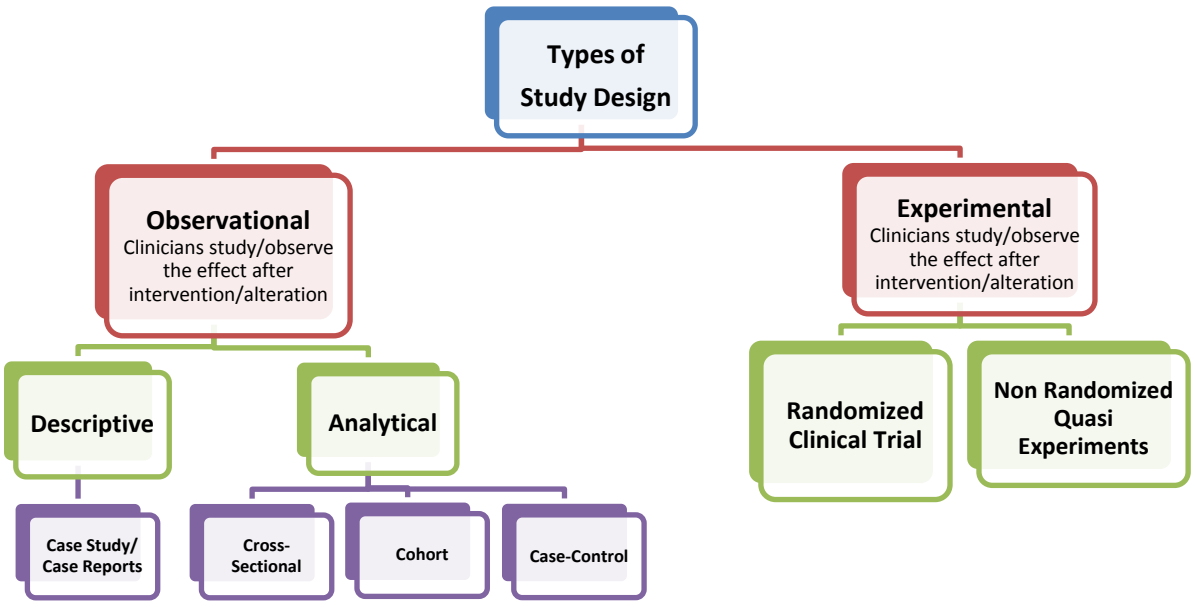
After framing the hypothesis, the study design is finalized.

1.3 Study Designs

Different study designs provide different types of information. This information is broadly classified as follows:

1. To assess the prevalence/incidence of a health issue
2. To identify the causative factors of a health issue
3. To determine the efficacy or safety of a treatment, drug, or therapy

Study designs can be categorized into the following types:



2. Hypothesis Testing

Statistical tests help to determine if one should accept or reject the null hypothesis. These tests determine the p-value associated with the research data. Assuming the null hypothesis (H_0) was true, the p-value is the probability that one could have obtained the result by chance.

2.1 Different Types of Hypothesis Tests

Different statistical tests can be applied on clinical data; however, all are based on the following steps:

1. State H_0
2. State H_1
3. Determine level of significance (α)
4. Determine the statistical tests to be applied
5. Determine the associated p-value
6. Accept or reject H_0

The statistical tests that are applied depend on the study design and the type of predictor and outcome variable (continuous/categorical). Few of the commonly used tests include:

1. T-test
2. Chi-squared test
3. Fisher's exact test
4. Log-rank test
5. Mann–Whitney test
6. Kruskal–Wallis test
7. ANOVA
8. Friedman test
9. Multivariate regression
10. Wilcoxon test and more

For instance, nominal variables can be tested using Chi-squared test or Fisher's exact test; continuous variables can be tested using T-test; or ordinal variables can be tested using Wilcoxon

rank Sum test or Mann–Whitney U test.

2.2 P-Value Approach

One must reject H_0 if the p-value of the data falls below the predetermined level of statistical significance (α). Usually, α is set at 0.05. If the p-value is less than 0.05, then H_0 is rejected, stating that there is no relationship between the predictor and the outcome in the sample population.

However, if the p-value is greater than α, then there is no statistically significant association between the predictor and the outcome variable. This does not mean that there is no association between the predictor and the outcome in the population; however, it indicates that the difference between the relationship observed and the relationship that could have occurred by random chance is small.

For instance, H_0 is the patients who take the drug that was used in the study after a heart attack did not have a better chance of not having a second heart attack over the next 24 months. Data suggests that those who did not take the drug that was used in the study were twice as likely to have a second heart attack with a p-value of 0.08. This p-value would indicate that there was an 8% chance that one would see a similar result (people on the placebo being twice as likely to have a second heart attack) in the general population because of random chance.

2.3 Misinterpretation of P-Value

Though p-value is an important statistical indicator, it has often been misinterpreted. For instance, a p-value of 0.05 does not mean that the finding is clinically relevant. A finding will have clinical significance only if the endpoint being studied actually has an impact on how a patient would be treated or diagnosed. Sometimes, it is impossible to directly study the variable that is relevant, so another variable or trait has to be used. A result involving this secondary trait may be of limited clinical value.

Another misconception is that observed data would occur only 5% of the time if the null hypothesis was true. In fact, the p-value is the probability that the observed data, plus more extreme data, would occur if the null hypothesis was true. More extreme data are usually unobserved and complicate the issue.

Some researchers believe that the p-value should be written as an inequality. This is based on the p-value serving as a cut off—the observed data is either statistically significant or not. However, reporting the absolute p-value of the hypothesis test indicates how much evidence is there to reject the null hypothesis. For instance, when

$p > 0.05$, results are not statistically significant

$0.05 \leq p < 0.1$, results can sometimes indicate a statistically significant observation

$0.01 \leq p < 0.05$, results are statistically significant

$0.001 \leq p < 0.01$, results are highly significant

$p < 0.001$, results are very highly significant

It is important to remember that a statistically significant result was initially meant to be an indication that an experiment was worth repeating. If the replication studies also yield statistically significant results, then the association is unlikely due to mere chance. The p-value should not be separated from the experimental data or its real world applications.

3. Types of Errors

Sometimes, the results obtained may not represent the population and lead to errors in making inferences. There are two types of errors.

3.1 Type I (False Positive)

It occurs if one rejects a null hypothesis when it is actually true in the study population. It is represented by α. For instance, H_0 is that the study drug has no effect on cholesterol level and α is 0.05 and H_1 is that the study drug has an effect on the cholesterol level.

If the statistical analysis revealed a p-value of 0.01, null hypothesis will be rejected. There is a 5% chance (α = 0.05) that the decision to reject the null hypothesis is wrong.

If H_0 is rejected, but in reality, the effect of the drug on the cholesterol level of the clinical trial participants is indistinguishable from random chance, then it is a type I error.

This means that one has incorrectly accepted H_1, but the drug actually has no real impact on the cholesterol level.

3.2 Type II (False Negative)

It occurs if one fails to reject a null hypothesis that is actually false in the study population. It is represented by β.

From the above example, if the p-value was 0.08, one would accept H_0. What if the drug has an effect on the cholesterol level? This would be a type II error implying that one has incorrectly concluded that the drug has no effect on the cholesterol level.

Hypothesis Testing	**H_0 is true**	**H_0 is false**
Reject H_0	Type I error (False positive)	True positive
Accept H_0	True negative	Type II error (False negative)

4. Power of a Statistical Test

The power of a statistical analysis is represented as $(1 - \beta)$. From the above example, if β is 0.1, it means that there is a 10% chance that one will not detect a relationship between the drug that is being studied and the cholesterol level, even though the drug has an effect on cholesterol level. In other words, there is a 90% chance of identifying a true relationship between the predictor and outcome variables in study population.

4.1 Effect Size

Represented by δ, effect size is the smallest difference that is clinically or biologically meaningful. This is the difference between the means (μ) of the two groups being compared. The effect size can be determined from the literature or by conducting a pilot study. A smaller δ will require a larger study population in order to detect or observe the effect. Determining the most appropriate effect size is thus crucial to calculate the size of the study population (sample size).

Effect size and power are also closely related. If effect size is large, for example, 75% of people whose parents had diabetes will develop diabetes themselves, then the true relationship between the variables will be easy to detect. This means that one can have a study with 90% or 95% power with a small sample size.

If the effect size is small, for example, 2% of the people on medication for cholesterol will hallucinate, one will need a very large study population in order to see the relationship between intake of this medication and hallucination. Usually, the effect size is unknown when initiating a clinical trial.

4.2 What Factors Affect Power?

Power is affected by following factors:

1. Sample size (n): Larger the sample size, smaller the power
2. Variance (σ2): Smaller the variance, larger the power
3. Study design
4. Effect size (δ): Larger the effect size, smaller the power
5. Level of significance (α): Larger the level of significance, larger the power
6. Type II error (β): Larger the type II error, smaller the power

The power of a clinical study should be minimum 80% and often study designs set power at 90-95% to observe the desired clinical effect. The 80% threshold represents a compromise between the likelihood to detect an effect, if one exists and the need of an incredibly large sample size.

5. Sample Size

One of the early steps in designing a clinical study is determining the sample size. If this calculation is not done correctly, it may result in a quantitative research that is unable to detect the true relationship between the predictor and outcome variables. The sample should be of appropriate size and representative of the population being studied. Sample size calculations require that the effect size, type I error level, type II error level, and the standard deviation (σ) are known.

5.1 What Factors Affect Sample Size?

1. Effect size (δ): Smaller the effect size, larger the sample size in order to detect or observe the effect.
2. Level of significance (α): Smaller the level of significance, larger the sample size.
3. Type II error (β): Smaller the type II error, larger the sample size in order to maintain the power of the study.
4. The standard deviation (σ): Greater the σ, larger the sample size required to attain the required power and level of significance.

Sample size calculations can be precision-based or power-based.

5.2 How to Calculate Sample Size?

Sample size tables and software programs are available to determine the appropriate sample size. However, it can be calculated using mathematical expressions depending on the study design, types of statistical tests used, and value of the parameters discussed above. Those mathematical expressions are not covered in this e-book. Attrition or dropout rate should also be factored in sample size calculation. For example, if the dropout rate is expected to be 20%, then the sample size should be increased by a factor of $1/(1 - 0.2)$, i.e., by 25%.

There can be limitations in calculating sample size as the mathematical expressions account for assumptions about effect size and some predefined parameters like type I & II errors. The sample size can also be constrained by implementation and administrative costs.

6. Relative Risk & Odds Ratio

In clinical research, risk is used to define an outcome and odds is used to define a precursor or antecedent. When comparison is made between two groups, the terms commonly used are relative risk and odds ratio. For instance, in the table below,

		Outcome	
Factor	Case group	Control group	Total
Yes	a	c	a+c
No	b	d	b+d
Total	a+b	c+d	a+b+c+d

6.1 Odds Ratio (OR)

Odds ratio is defined as the odds of a disease in case group to that in the control group. For instance, a study observes occurrence of total hip replacement in two groups- one was diagnosed with osteoporosis earlier and the other was not diagnosed. If the OR for this study is 2/5 (0.4), it implies that the diagnosis of osteoporosis reduced the probability of total hip replacement by 0.4%. Mathematically, RR can be expressed as follows:

$$\mathrm{OR} = \frac{a/b}{c/d}$$

6.2 Relative Risk (RR)

Relative risk is defined as the ratio of risk of an outcome in case group to that in the control group. For instance, if RR of having a heart disease in a group with hypertension to those without hypertension is 10.5, it implies that first group is 10.5 times likely to have a heart disease. Mathematically, RR can be expressed as follows:

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

7. Correlation & Regression

Clinicians may not only wish to know the variables associated with a disease or condition but also the relationship and strength of relationship among those variables. Correlation and regression analysis thus help quantifying that relationship.

7.1 Regression Analysis

A linear regression involves one independent variable and the outcome variable. This should be used to model a relationship between data if the dependent variable is continuous. A multiple regression involves two or more independent variables that are expected to influence the outcome variable. A logistic regression would be used to model data if the dependent variable is dichotomous. In each case, analysis is performed to model any statistical relationship between the dependent and independent variables.

For instance, asthmatic patients who live in polluted areas and have smokers in their homes use more asthma medication than those who do not. A multiple regression analysis could be used to find out if there is an actual association between smokers in the household and air pollution and how often patients have an asthma attack. In this case, the number of inhalers purchased in a time period would be used as the dependent variable. The independent variables that are influencing this outcome would be air pollution and presence of a smoker in the house.

Required data would be collected and a statistical program, such as SPSS or R would be used to plot a relationship among these three variables. The program would draw a regression line. The regression line approximates the relationship among the variables. The statistical program will also give you a formula that explains the relationship. Usually, it would have the format

$$\text{Linear regression, } Y = mX + c + \text{error term}$$

$$\text{Multiple regression, } Y = m_1X_1 + m_2X_2 + m_3X_3 + c + \text{error term}$$

Y is the dependent variable (number of inhalers purchased in a year). X is the independent variable, and c is a constant. The error term indicates how confidently the relationship can be predicted. The smaller the error term, the more certain one can be of the regression line. “m” is a factor that indicates the influence of X. For instance, if m = 10 and X_1 represented the impact of

smoking and X_2 represented the impact of air pollution on triggering asthma attacks, one could say that for patients who live in an area with very clean air and no smokers in the family, the number of inhalers purchased in a year would be c. For every unit increase in air pollution and number of smokers the asthmatic patient is exposed to the number of inhalers purchased in a year will increase by 10.

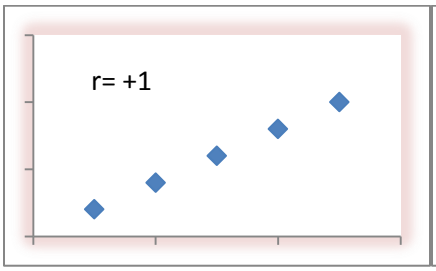
7.2 Correlation

When the regression equation is determined, the correlation coefficient (r) predicts the nature and strength of the observed relationship. Value of r may vary from +1 to −1. If r = 0, it implies that there was no relationship observed. Different correlation coefficients like Pearson, Kendall, and Spearman can be calculated using statistical softwares based on types of variables.

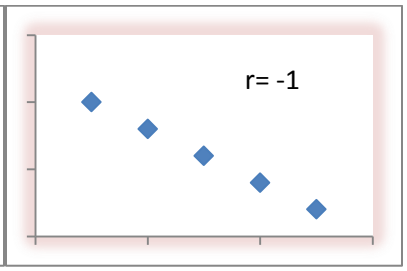
7.3 Types of Correlation

A correlation between variables can be positive, negative or zero as represented by scatter point graph.

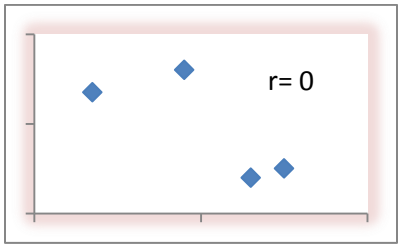
Positive Correlation



Negative Correlation



Zero Correlation



Note: A zero correlation does not always indicate absence of relationship; it can also mean that the relationship is linear.

The measure r^2 is a fraction of the variation in dependent variable that is explained by the regression model. In other words, the r^2 value associated with your regression analysis will indicate how much of the relationship has been explained by the regression model. An r^2 value of 52% for this asthma study (considering only independent variable i.e. impact of air pollution) would indicate that air pollution explain 52% of the variability in the number of inhalers asthmatics buy. This would suggest other factors that need to be included to explain a larger proportion of the variation observed.

7.4 Note of Caution

If you use regression analyses to identify a relationship among variables, you should not stop there. Ask, "Does this relationship make sense?" One may need to design additional experiments to test the regression relationship in the real world. It is also critical to remember that *correlation does not equal causation*. For instance, you may find that more people drown during a heat wave. It would be incorrect to say that heat waves cause drowning. Probing the relationship some more might reveal that more people go to the beach during heat waves, increasing the likelihood of drowning.

8. Multiplicity Issue

Clinical trials very often involve performing multiple comparisons on the clinical data. This leads to multiplicity problem. It occurs because the more statistical tests one runs, the higher the likelihood of achieving the statistical significance. If the level of significance is 0.05, the probability of getting at least one statistically significant result is 64% if 20 tests are done and rises to 99.4% if 100 tests are done. The problem with multiple testing is that it is highly likely to obtain a p-value below the level of significance even when there is no true association. There are guidelines on when statistical corrections are required to deal with the multiplicity problem and when these corrections are not necessary. Clinical trials that are free from this multiplicity issue have following characteristics:

1. Only two treatment groups
2. Utilize one primary variable
3. Have a strategy in place to confirm the results using a single null hypothesis involving the primary variable, which lacks any interim analysis

Multiple testing occurs when more than one independent statistical test is performed on the data. This can occur in any of the clinical trial phases.

For instance, multiple testing would occur if a medicine in a clinical trial targets more than one symptom. Relief from each symptom would be considered an endpoint. It is, therefore, possible that the drug being tested has a statistically significant effect on patient symptoms just because multiple statistical tests were applied. This problem can also be introduced if there are many subgroups in a clinical trial.

8.1 Multiplicity Adjustment

Multiple testing can be done without introducing false associations if certain statistical measures or corrections are applied. The **Bonferroni correction (α/n)** is one of those measures. The Bonferroni correction adjusts the significance level (α) by dividing the cut-off point (usually

0.05) by the number of independent tests being done (n). This correction is known to be conservative and can lead to high rate of false negatives.

Bioinformatics, especially genome-wide association studies, frequently perform thousands of independent tests on the same data set. In this case, **controlling the false discovery rate (FDR)** might be the better correction to use. The FDR determines the proportion of false positives (type I error) among all the results deemed significant. The FDR should be less than the level of significance (α). This correction, thus, helps to control the type I errors.

8.2 Clinical Trial Adjustments

Another approach that researchers can take when performing medical research or clinical trials is to choose a single primary endpoint. As an example, a cold medicine might have desired endpoints of reducing congestion, lowering fever, and alleviating post-nasal drip. Instead of determining if the drug has a statistically significant effect on all three symptoms, alleviating congestion could be made the primary endpoint for biostatistical calculations.

If any of the clinical trial phases involves multiple subgroups, the statistical tests can still be performed. The level of significance should not be altered after the trial has been completed, the number of subgroups should be kept low, and the results should be biologically plausible and aligned with the external evidence. If these conditions are not met, the results should be viewed as preliminary and another clinical trial should be planned to closely examine other statistically significant results.

The multiplicity curse may be the reason that Phase III clinical trials fail due to many endpoints and comparisons (Phase III trials are supposed to confirm the safety and efficiency data derived from Phase II trials). This may arise due to many reasons. Repeated measures during the clinical trial can be one of them. This might happen if the study requires measurements of the same patient over a given time period and can be avoided by using a summary measure such as the mean or median of all the readings. One could also reduce the number of time points in the clinical trial.

Important Global Events

Harvard Medical School – Research Misconduct Leads to $10 Million Settlement

https://www.enago.com/academy/harvard-medical-school-research-misconduct-leads-to-10-million-dollars-settlement/

Chinese Scientists Increase Global Research Collaboration

https://www.enago.com/academy/chinese-scientists-increase-global-research-collaboration/

6 Tips That Will Save You From Journal Rejection

https://www.enago.com/academy/6-tips-that-will-save-you-from-journal-rejection/

Measuring Research Impact: Citation Metrics

https://www.enago.com/academy/measuring-research-impact-citation-metrics/

Some Important Links

Latest News

https://www.enago.com/academy/category/industry-news/publishing-hot-topics/

Featured Interviews

https://www.enago.com/academy/category/featured-interviews/

Manuscript Preparation Resources

https://www.enago.com/academy/category/academic-writing/